Deinterleaving Markov Processes via Penalized ML

Gadiel Seroussi
Hewlett-Packard Laboratories
Palo Alto, CA, USA
Email: gseroussi@ieee.org

Wojciech Szpankowski
Department of Computer Science
Purdue University
West Lafayette, IN, USA
Email: spa@cs.purdue.edu

Marcelo J. Weinberger
Hewlett-Packard Laboratories
Palo Alto, CA, USA
Email: marcelo.weinberger@hp.com

***Abstract*—We study the problem of deinterleaving a set of finite memory (Markov) processes over disjoint finite alphabets, which have been randomly interleaved by a memoryless random switch. The deinterleaver has access to a sample of the resulting interleaved process, but no knowledge of the number or structure of the Markov processes, or the parameters of the switch. We present a deinterleaving scheme based on minimizing a penalized maximum-likelihood cost function, and show it to be strongly consistent, in the sense of reconstructing, almost surely as the observed sequence length tends to infinity, the original Markov and switch processes. Solutions are described for the case where a bound on the order of the Markov processes is available, and for the case where it is not. We demonstrate that the proposed scheme performs well in practice, requiring much shorter input sequences for reliable deinterleaving than previous solutions.**

I. Introduction

Problems in applications such as data mining, computer security, and genomics, often require the identification of streams of data from different sources, which may be intermingled or hidden (sometimes purposely) among other unrelated streams. The source identification problem studied in this paper is motivated by these problems (more detailed descriptions of the applications can be found in [1], [2]).

Consider m independent finite-memory (Markov) sources $P_1, P_2, \ldots, P_m$, defined over disjoint finite alphabets $A_1, A_2, \ldots, A_m$, respectively, and of order at most k, i.e., for a sequence $x^t = x_1\, x_2 \ldots x_t \in (A_i)^t$, $t > k$, emitted by P_i, we have $P_i(x_t|x^{t-1}) = P_i(x_t|x_{t-k}^{t-1})$, $1 \le i \le m$.[1] Consider also a distribution $P_{\rm w}$ over $\{1, 2, \ldots, m\}$, and a memoryless random switch, which, at time t, selects one of the sources P_i with probability $P_{\rm w}(i)$, independently of past selections or source samples, and outputs the next sample from P_i. We assume that the sources P_i are idle, i.e., they do not emit symbols and they preserve their state, except when selected by the switch. We also assume that $P_{\rm w}(i) > 0$ and that $P_i(a) > 0$ for all $a \in A_i$, $1 \le i \le m$. Let $A = A_1 \cup A_2 \cup \cdots \cup A_m$, and denote by Π the partition $\{A_1, A_2, \ldots, A_m\}$ of A. The output of the switch is a process P over A, which can be seen as a random interleaving of the processes $P_1, P_2, \ldots, P_m$, governed by the switch $P_{\rm w}$. We refer to P as an *interleaving of Markov processes* (IMP), write $P = \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_{\rm w})$, and refer to this expression as an *IMP presentation* of P. To distinguish the alphabet A from its subsets A_i, we will refer to the latter as *sub-alphabets*.

Given a sample z^n from P, the problem of interest is to reconstruct the alphabet partition Π, and, consequently, the original sequences from the Markov processes, and the sequence of switch selections. Notice that, given P, any partition Π' of A induces a representation of P as an interleaving of processes, governed by a random switch. In general, if Π' is the "wrong" partition, the induced sub-processes P'_i will not be of finite order, the switch process will not be memoryless, or some of the independence assumptions will be violated. In some cases, however, different partitions might induce different valid IMP presentations of P. In such cases, we require that all the valid solutions be identified. We assume that the alphabet A is known, but no nontrivial knowledge of the structure or number m of the sub-alphabets A_i is assumed. We will consider both the case where the maximal Markov order k is known, and the case where it is not.

In this extended summary, we focus on a version of the problem where the switch is memoryless and the source sub-alphabets are disjoint. The case of a switch with memory can be solved using tools similar to the ones presented here, and will be treated in the full version of the paper. A feasible solution for the case of intersecting sub-alphabets appears more challenging, and is the subject of further research. Notice that in this case, identifying the sub-alphabets is not sufficient to deinterleave the constituent data streams.

The deinterleaving problem for the case $k = 1$ was studied in [1], where an approach was proposed that could identify a valid IMP presentation of P with high probability as $n \to \infty$ (the approach as described cannot identify multiple valid solutions when they exist). The idea is to run a greedy sequence of tests, checking equalities and inequalities between various event probabilities (e.g., $P(ab) \neq P(a)P(b)$, $P(abc) = P(a)P(b)P(c)$, $a, b, c \in A$), and permanently clustering symbols into sub-alphabets sequentially, according to the test results. Empirical distributions are used as proxies for the true ones. Clearly, equalities between probabilities translate only to "approximate equalities" subject to statistical fluctuations in the corresponding empirical quantities, and an appropriate choice of the tolerances used, as functions of the input length n, is crucial to turn the conceptual scheme into an effective

[1] We use standard notations x^n, x_i^j, etc., for sequences. 'Source' and 'process' are used interchangeably, preferring 'source' when emphasizing the generation mechanism, and 'process' when emphasizing the random sequence. Markov processes are assumed ergodic, but not necessarily stationary, as we assume fixed initial states. Marginal probability notations (e.g., $P_i(u)$ for a string u) represent the steady-state probabilities of their arguments.

algorithm. Specific choices for tolerances are not discussed in [1]. The attractive feature of the approach in [1] is its low complexity; equipped with a reasonable choice of tolerance thresholds, a very efficient algorithm for $k = 1$ can be implemented. More recently, the problem of deinterleaving hidden-Markov processes was studied, mostly experimentally, in [2].

In this paper, we present deinterleaving schemes that identify, eventually almost surely, all the valid IMP presentations of the observed process. The solution is based on finding the partition Π of A that minimizes a *penalized maximum likelihood* (penalized ML) cost function of the form $C(z^n, \Pi) = n\hat{H} + \beta K \log n$, where $\hat{H}$ is the empirical entropy of the observed sequence under an IMP model induced by Π, K is the total number of parameters in the model, and β is a nonnegative constant. Penalized ML estimators of Markov process order are well known (cf. [3], [4], [5]). Here, we use them to estimate the original partition Π, and, in the case of unknown k, also the Markov order of the processes P_i. When k is known, $\beta = 0$ (i.e., using a plain ML estimator) suffices in cases where none of the processes P_i is memoryless; an appropriate $\beta{>}0$ is necessary to identify all the IMP presentations when there are memoryless components P_i. In addition, an appropriate value $\beta{>}0$ is needed in the unknown k case, with the traditional function of penalizing larger models. In all cases, our estimate of the correct partition Π is strongly consistent.

The rest of the paper is organized as follows. In Section II we present the IMP P as a *finite state machine* (FSM) process. We also fully characterize the cases where the IMP presentation of P is not unique, which can occur only when some of the constituent Markov processes are memoryless. When $k{=}1$, our characterization reduces to that of [1]. In Section III we study the case of arbitrary but known k, and in Section IV that of unknown k. Finally, in Section V, we show results for practical implementations of our scheme and that of [1] (with optimized tolerances), for the case $k{=}1$. The results show that in addition to its asymptotic properties, the proposed deinterleaver achieves high accuracy rates in identifying the correct alphabet partition for much shorter sequences than those required by the scheme of [1]. The latter has the attractive feature of low computational complexity, while our ideal scheme calls for finding the optimal partition through an exhaustive search, which is computationally expensive. As we show in Section V, by sacrificing the optimality guarantee, and implementing a randomized gradient descent heuristic that searches for the same optimal partition, we obtain, in practice, the same results as with exhaustive search, but with a much faster and practical scheme.

II. FSM sources and IMP presentations

An FSM over an alphabet $\mathcal{A}$ is defined by a triplet $F = (S, s_0, f)$, where S is a set of *states*, $s_0 \in S$ is an *initial state*, and $f : S \times \mathcal{A} \to S$ is a *next-state function*. A (unifilar) *FSM source* is defined by associating a conditional probability distribution $P_F(\cdot|s)$ with each state of S. To generate a random sequence x^n, the source draws, for each i, $1 \le i \le n$, a symbol $x_i{\in}\mathcal{A}$ distributed according to $P_F(\cdot|s_{i-1})$, and transitions to the state $s_i = f(s_{i-1}, x_i)$. Markov sources of order k over $\mathcal{A}$ are special cases of FSM sources with $S = \mathcal{A}^k$. We next observe that an IMP can be represented as an FSM source.

Let $P = \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_{\rm w})$, and let k denote an upper bound on the Markov orders $\mathrm{ord}(P_i)$ of the processes P_i. For $a{\in}A$, let $\sigma_\Pi(a)$ denote the index of the sub-alphabet containing a, i.e., $a{\in}A_{\sigma_\Pi(a)}$. Denote by $S_i = (A_i)^k$ the set of states of P_i. We define an FSM $\mathcal{F}_k(\Pi) = (S, \mathbf{s}_0, f)$ with $S = S_1{\times}S_2{\times}\cdots{\times}S_m$, $\mathbf{s}_0 = (s_0^{(1)}, s_0^{(2)}, \ldots, s_0^{(m)})$ where $s_0^{(i)}$ denotes an initial state for P_i, and f a next-state function defined as follows: For $a \in A_i \subseteq A$, let f_i denote the next-state function of P_i. Given a state $\mathbf{s} = (s^{(1)}, s^{(2)}, \ldots, s^{(m)}) \in S$, we define $f(\mathbf{s}, a) = \mathbf{s}' = (s'^{(1)}, s'^{(2)}, \ldots, s'^{(m)})$, where $s'^{(j)} = s^{(j)}$ for $j \ne i$, and $s'^{(i)} = f_i(s_i, a)$. Similarly, for each state $\mathbf{s} \in S$, we define a conditional probability distribution

$$P_{\Pi,k}(a \,|\, \mathbf{s}) = P_{\rm w}(i) P_i(a \,|\, s^{(i)}), \;\; i = \sigma_\Pi(a), \; a \in A\,. \qquad (1)$$

The following proposition is readily verified.

Proposition 1: $\mathcal{F}_k(\Pi)$, with transition probabilities $P_{\Pi,k}$, generates P.

Let $\alpha_i = |A_i|$, $1 \le i \le m$, and $\alpha = |A|$. An arbitrary FSM source over A, with underlying FSM $\mathcal{F}_k(\Pi)$, has, in general $D = (\alpha - 1)\prod_{i=1}^m \alpha_i^k$ free statistical parameters. The conditional probability distributions in (1), however, are highly constrained, as the parameter $P_{\Pi,k}(a|\mathbf{s})$ is constrained to have the same value for all states $\mathbf{s}$ with the same value of $s^{(i)}$, where $i = \sigma_\Pi(a)$. Overall, the number of free parameters remains, of course, that of the original Markov processes and switch, i.e., $d = \sum_{i=1}^m \alpha_i^k(\alpha_i - 1) + (m - 1)$. We refer to sources satisfying the constraints implicit in (1) as *IMP-constrained* FSMs (in short, IFSMs). Let V denote the set of parameter vectors corresponding to valid unconstrained FSMs based on $\mathcal{F}_k(\Pi)$. This set is a convex polytope in D-dimensional Euclidean space. Since the IMP constraints are equalities, and the multiplications by the quantities $P_{\rm w}(i)$ (which are parameters) can be translated to boundary conditions, the parameter vectors of IFSMs based on $\mathcal{F}_k(\Pi)$ lie, within V, on a d-dimensional hyperplane. We shall make use of this geometry in Section III.

Notice that, in general, P is not a finite memory process. One special case where P does have finite memory is when all the P_i are memoryless, and, thus, so is P. Clearly, if $m > 1$ and any of the processes P_i has memory, then P cannot have finite memory. Thus, either P is memoryless or it has infinite memory. The memoryless case is special also because unless $|A|{=}1$, a memoryless process P over A admits multiple IMP presentations. For example, if P is memoryless, the partitions $\{\{a_1\}, \{a_2\}, \ldots, \{a_m\}\}$, $m > 1$, with $P_{\rm w} = P$, and $\{\{A\}\}$, with $P_{\rm w}(1) = 1$, induce valid IMP presentations of P. In fact, it can readily be verified that any partition of A defines a valid IMP presentation in this case. With a slight abuse of terminology, we say that $A_i \in \Pi$ is a *memoryless sub-alphabet* if P_i is memoryless, the process setting being understood from the context. In the remainder of the section, we show that the IMP presentation of a process is unique up to re-partitioning

of memoryless sub-alphabets.

We say that Π is *canonical* with respect to P if all its memoryless sub-alphabets are singletons. By the discussion above the partition Π in an IMP can always be brought to canonical form, without changing the process, by splitting its memoryless sub-alphabets into singletons. We denote the resulting partition by Π^* (dependence on P being understood from the context). We say that a partition Π' is *compatible* with P if there exist processes $P'_1, P'_2, \ldots, P'_{m'}$ and $P'_{\rm w}$ such that $P = \mathcal{I}_{\Pi'}(P'_1, P'_2, \ldots, P'_{m'}; P'_{\rm w})$. We say that two partitions Π, Π' are *equivalent* with respect to P, denoted $\Pi \equiv_P \Pi'$, if they are both compatible with P, and $\Pi^* = \Pi'^*$. For an alphabet $\mathcal{A}$, a subset $\mathcal{A}' \subseteq \mathcal{A}$, and a sequence $u^t \in \mathcal{A}^t$, we denote by $u^t[\mathcal{A}']$ the sub-sequence of u^t composed of its symbols in $\mathcal{A}'$.

Theorem 1: Let P be an IMP, and let Π be a partition of A compatible with P. A partition Π' of A is compatible with P if and only if $\Pi \equiv_P \Pi'$.

Proof: The 'if' part follows from the discussion above. Assume $P = \mathcal{I}_{\Pi'}(P'_1, P'_2, \ldots, P'_{m'}; P'_{\rm w})$. To prove the 'only if' part, it suffices to show that if symbols that are in the same sub-alphabet in one partition become separated in the other, these symbols must be associated with memoryless processes in both presentations of P. Assume $A_i \in \Pi$ and $A'_j \in \Pi'$ are such that $A_i \cap A'_j \neq \phi$, and $A_i \setminus A'_j \neq \phi$. We first claim that for all $t > 0$, there exists a string $u^t \in (A_i \setminus A'_j)^t$ such that $P(u^t) > 0$. First, from the definitions of A_i and A'_j, there exists a symbol $b \in A_i \setminus A'_j$, and by our basic assumptions, $P(b) > 0$. This proves the claim for $t = 1$. Assume now that a string u^τ satisfies the claim for some $\tau > 0$. We have

$$\sum_{x \in A_i \cap A'_j} P(x|u^\tau) = \sum_{x \in A_i \cap A'_j} P(x) = P(A_i \cap A'_j) < P(A_i)\,,$$

where the first equality follows from the fact that in the presentation $\mathcal{I}_{\Pi'}$, symbols in A'_j are independent of those in $A_i \setminus A'_j$. Therefore, there must exist a symbol $c \in A_i \setminus A'_j$ such that $P(c|u^\tau) > 0$, and, thus, $P(u^\tau c) > 0$, which establishes the auxiliary claim.

Now, let $a \in A_i \cap A'_j$, let $s \in A_i^k$ with $P_i(s) > 0$, where k is the order of P_i, and let $s' = s[A'_j]$. Clearly, $P(s) > 0$ and $P(s') > 0$, and we have

$$P_i(a|s) = P(a|s)/P_{\rm w}(i) = P(a|s')/P_{\rm w}(i), \qquad (2)$$

where the first equality follows from the definition of $\mathcal{I}_\Pi$ while the second follows from that of $\mathcal{I}_{\Pi'}$. Choose a string u^k satisfying the auxiliary claim. Since, in $\mathcal{I}_{\Pi'}$, s' is independent of u^k, we have $P(s'u^k) > 0$. Now, by the properties of $\mathcal{I}_\Pi$, and the definition of k, we have $P(a|s') = P(a|s'u^k) = P(a|u^k)$, independent of s'. Recalling (2), we conclude that $P_i(a|s)$ is independent of s. Since A'_j can be any sub-alphabet in Π' that intersects A_i, the statement is true for all $a \in A_i$ and state s of P_i. Hence, P_i is memoryless as claimed. Clearly, P'_j is memoryless if A'_j is a subset of A_i. If it is not, we reverse the roles of Π and Π', and let A'_j play the role of A_i in the proof. Thus, P'_j is memoryless. ∎

III. Solution for known k

For a sequence x^t, and an integer $k \geq 0$, denote by $\hat{H}_k(x^t)$ the kth order (per-symbol) empirical entropy of x^t,

$$\hat{H}_k(x^t) = -\frac{1}{t} \log P_k^{\rm ML}(x^t),$$

where $P_k^{\rm ML}(x^t)$ is the ML (or empirical) probability of x^t under a kth order Markov model with a fixed initial state. Consider now a sequence z^n over A, and a partition $\Pi = \{A_1, A_2, \ldots, A_m\}$ of A. We extend the definition of the function σ_Π to sequences, defining $\sigma_\Pi(z^n) = \sigma_\Pi(z_1), \sigma_\Pi(z_2), \ldots, \sigma_\Pi(z_m)$, where we recall that $\sigma_\Pi(z_i)$ is the integer satisfying $z_i \in A_{\sigma_\Pi(z_i)}$. The sequence $\sigma_\Pi(z^n) \in \{1, 2, \ldots, m\}^n$ can be regarded as a "switch sequence" induced by z^n and the partition Π. Let $n_i = \big|z^n[A_i]\big|$, $1 \leq i \leq m$. For $F = \mathcal{F}_k(\Pi)$, define

$$\hat{H}_F(z^n) \;= \sum_{i=0}^{m} \frac{n_i}{n} \hat{H}_k(z^n[A_i]) \;+\; \hat{H}_0(\sigma_\Pi(z^n)), \qquad (3)$$

$$L_F(z^n) \;= \beta \sum_{i=1}^{m} (\alpha_i - 1)\alpha_i^k \log n + \beta(m-1)\log n\,, \quad (4)$$

where β is a nonnegative constant. $\hat{H}_F(z^n)$ can be regarded as the empirical entropy of z^n with respect to F *as an IMP-constrained FSM* (which is different from its empirical entropy with respect to F as an *unconstrained* FSM, where no equalities between statistics are imposed). $L_F(z^n)$, in turn, is a penalty term proportional to the total number of parameters in the constrained model. We will refer to the first term on the right-hand side of both (3) and (4) as "process contribution," and to the second term as "switch contribution." We now define the penalized ML *cost function*

$$C_F(z^n) = n\hat{H}_F(z^n) + L_F(z^n)\,. \qquad (5)$$

Our deinterleaving scheme selects its solution $\hat{F}_k(z^n)$ according to the following rule:

$$\hat{F}_k(z^n) = \arg \min_{F' = \mathcal{F}_k(\Pi')} C_{F'}(z^n)\,, \qquad (6)$$

where Π' ranges over all partitions of A.

Theorem 2: Let $P = \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_{\rm w})$, let Π^* denote the associated canonical partition, and let z^n be a sample from P. Then, for appropriate choices of β in (4) (e.g., $\beta = \frac{1}{2}$), we have

$$\hat{F}_k(z^n) = \mathcal{F}_k(\Pi^*) \quad \text{a.s. as } n \to \infty\,.$$

Proof outline: Consider a partition Π' such that $\Pi \not\equiv_P \Pi'$, and let $\mathcal{F}_k(\Pi')$ be the associated FSM. Let F^+ be a common refinement[2] of $F = \mathcal{F}_k(\Pi)$ and $F' = \mathcal{F}_k(\Pi')$. With appropriate parameter settings, F^+ can generate all the processes that either F or F' can generate. In particular, $\hat{P}_F(z^n) = 2^{-n\hat{H}_F(z^n)}$ and $\hat{P}_{F'}(z^n) = 2^{-n\hat{H}_{F'}(z^n)}$ are prob-

[2] A *refinement* [6] of $F = (S, s_0, f)$ is an FSM $F^+ = (S^+, s_0^+, f^+)$ such that for some fixed function $g : S^+ \to S$ and any sequence x^n, the respective state sequences $\{s_i\}$ and $\{s_i^+\}$ satisfy $s_i = g(s_i^+)$, $0 \leq i \leq n$. A refinement of F can generate all the processes that F generates. It is always possible to construct a common refinement of two FSMs.

abilities that can be assigned by F^+ to z^n. Let $P^{\mathrm{ML}}_{F^+}(z^n)$ denote the ML probability of z^n with respect to F^+ as an *unconstrained* FSM. Then, by the definitions of $P^{\mathrm{ML}}_{F^+}$ and $D(\cdot||\cdot)$, the Kullback-Liebler divergence between FSM processes, we have

$$n\hat{H}_{F'}(z^n) = -\log P^{\mathrm{ML}}_{F^+}(z^n) + nD(P^{\mathrm{ML}}_{F^+}||\hat{P}_{F'})\,, \quad (7)$$

$$n\hat{H}_{F}(z^n) = -\log P^{\mathrm{ML}}_{F^+}(z^n) + nD(P^{\mathrm{ML}}_{F^+}||\hat{P}_{F})\,. \quad (8)$$

Let V denote the space of all valid parameter vectors for FSM sources based on F^+. The constraints satisfied by the IMP sources based on $\mathcal{F}_k(\Pi)$ and $\mathcal{F}_k(\Pi')$ are inherited by their representations in V. Thus, as discussed in Section II, the set of all IFSM sources based on $\mathcal{F}_k(\Pi')$ lies, within V, on a hyperplane $V'\subseteq V$ (notice that the additional constraints resulting from the refinement are also in the form of parameter equalities). Since, by Theorem 1, no IMP assignment of parameters for $\mathcal{F}_k(\Pi')$ can generate P, the representation of P in[3] V is outside of V', and, hence, at positive Euclidean (or L_1) distance from it. Thus, by Pinsker's inequality applied state by state in F^+, for any process P' in V', we have

$$D(P||P') \geq \Delta > 0\,, \quad (9)$$

for some positive constant Δ. Since $\hat{P}'_{F'}$ is in V', and the parameters of both $P^{\mathrm{ML}}_{F^+}(z^n)$ and $\hat{P}_F(z^n)$ converge to the true parameters of P a.s., it follows from (7)–(9) that

$$\hat{H}_{F'}(z^n) - \hat{H}_F(z^n) \geq \Delta\,, \text{ a.s. as } n\to\infty\,, \quad (10)$$

which implies, by (5),

$$n^{-1}C_{F'}(z^n) - n^{-1}C_F(z^n) \geq \Delta\,, \text{ a.s. as } n\to\infty\,, \quad (11)$$

since the contribution of the $O(\log n)$ penalty terms to the costs vanishes asymptotically in this case. It follows that when comparing Π with a non-equivalent partition, the rule (6) will select Π^* eventually almost surely, for any value of β. In particular, it will do so for $\beta = 0$, i.e., a plain ML estimator. We now discuss the case where partitions that are equivalent to Π are compared. By Theorem 1, these partitions can differ only by a re-partitioning of memoryless sub-alphabets. Consider a partition $\Pi' \equiv_P \Pi$, assume $A'_i \in \Pi'$ is memoryless, and let $\alpha'_i = |A'_i| > 1$. Let Π'' be the partition derived from Π' by splitting A'_i into singletons, and let $F'' = \mathcal{F}_k(\Pi'')$. Singleton sub-alphabets contribute zero cost to the "process" terms of (3) and (4); their contribution to the overall cost is contained in the "switch" terms. Let $u^{n_i} = z^n[A'_i]$. By carrying out the computation of cost shifts from "process" terms to "switch" terms when A'_i is split, we obtain

$$\begin{aligned} C_{F'}(z^n) - C_{F''}(z^n) = \quad (12)\\ n_i\hat{H}_k(u^{n_i}) + \beta{\alpha'_i}^k(\alpha'_i-1)\log n \\ -\big(\, n_i\hat{H}_0(u^{n_i}) + \beta(\alpha'_i-1)\log n\,\big)\,. \end{aligned}$$

The difference in (12) can be seen as the difference in costs that would result from comparing order k with order 0 with a suitably penalized ML order estimator for u^{n_i} (having $\log n$ instead of $\log n_i$ in (12) does not affect the conclusions). Since u^{n_i} is a sample of a memoryless process, it follows from the consistency of these estimators (cf. [5]) that the difference must be positive eventually a.s. Thus, lowest cost is attained by splitting all memoryless sub-alphabets into singletons, and (6) will return $\mathcal{F}_k(\Pi^*)$ eventually a.s. ■

By Theorem 2, if P has memoryless components, the rule (6) will return the symbols in these components as a list of singletons. From this list, it is possible to reconstruct all the partitions that are equivalent to Π. Since the empirical entropy decreases with the order, a plain ML estimator, on the other hand, would have returned all memoryless sub-alphabets merged into one, which without further processing would not be recognized as memoryless (as is the case in [1]).

IV. Solution for unknown k

Penalized ML Markov order estimators are known to be strongly consistent also when a bound on the order is not known [7], [5], and even when the order is allowed to grow with the sample length n, as long as the growth rate is slow enough [8]. Thus, it would be natural to expect that such an estimator might be useful also for deinterleaving IMPs in cases where the maximum order of the constituent processes is unknown. The application is not straightforward, though, since in most cases, when the partition being tested is incorrect, the cost function (5) is evaluated on samples from processes that do not have finite memory, nor an order that can be bounded by a slowly growing function of n. In this section, we prove the strong consistency of a penalized ML estimator for IFSMs when a finite bound on the order of the Markov processes P_i exists but is unknown to the deinterleaver. The proof borrows some elements from [9], where a strongly consistent estimator for unconstrained FSMs was presented.

We define the estimator

$$\hat{F}(z^n) = \arg\min_{F'=\mathcal{F}_{k'}(\Pi')} C_{F'}(z^n)\,, \quad (13)$$

where $C_{F'}$ is as defined in (5), and the minimization is over all partitions Π' and orders $k' \geq 0$. Notice that the number of parameters of the minimizing FSM will never exceed n (a memoryless model would attain a lower cost than a model with n parameters), so k' can be limited in value to $O(\log n)$, and the search is finite.

Theorem 3: Let $P = \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_{\mathrm{w}})$, and let $k = \min_{k'}\{\, k' \,|\, \mathrm{ord}(P_i) \leq k',\, 1 \leq i \leq m\}$, the minimal upper bound on the orders of the processes P_i. Let Π^* be the canonical partition associated with P. Then, for a sample z^n of P, using $\beta > 2$ in (4)–(5), we have

$$\hat{F}(z^n) = \mathcal{F}_k(\Pi^*) \quad \text{a.s. as } n\to\infty\,.$$

Proof outline: Let $F{=}\mathcal{F}_k(\Pi^*)$, and let $F'{=}\mathcal{F}_{k'}(\Pi')$ be an FSM in the search space of (13). Let $\mathcal{P}_{F'}$ denote the probability that F' be selected in the minimization. We bound the probability of selecting the wrong FSM in (13) by summing $\mathcal{P}_{F'}$ over all $F' \neq F$. Let $Q(F')$ denote the number

[3]Here we assume, for simplicity, that the representation of P as a point in V is unique, which is the case if $P_i(s) > 0$ for all states s in all constituent processes P_i. The restriction is easily removed through a slightly lengthier argument, which will be given in the full paper.

of free parameters in F', let $K = Q(F)$, and let $Q_0 > K$ be a threshold for $Q(F')$, which is independent of n, and will be specified later. As in [9], we divide the error probability in two parts, as follows:

$$\sum_{F' \neq F} \mathcal{P}_{F'} = \sum_{Q(F')<Q_0,\ F' \neq F} \mathcal{P}_{F'} + \sum_{Q(F') \geq Q_0} \mathcal{P}_{F'} \,. \quad (14)$$

The FSMs covered by the first summation involve Markov processes of bounded order. This case is handled by a straightforward modification (to be described in the full paper) of the arguments in the proof of Theorem 2. Here, we focus on the second term, which requires different tools. Let $B_{F'} = \{z^n \,|\, C_{F'}(z^n) \leq C_F(z^n)\}$. We have

$$\mathcal{P}_{F'} = \sum_{z^n : \hat{F}(z^n) = F'} P(z^n) \leq \sum_{z^n \in B_{F'}} P(z^n). \quad (15)$$

From the definition of empirical entropy, (5), and the definition of $B_{F'}$, we have, for $z^n \in B_{F'}$,

$$P(z^n) \leq 2^{-n\hat{H}_F(z^n)} \leq 2^{-n\hat{H}_{F'}(z^n) - L_{F'}(z^n) + L_F(z^n)} \,. \quad (16)$$

From (16) and (4), setting $q = Q(F')$, we get

$$P(z^n) \leq \, 2^{-n\hat{H}_{F'}(z^n)} \cdot n^{\beta(K-q)} \,, \quad z^n \in B_{F'} \,.$$

Thus, by (15), but summing over all of A^n, we get

$$\mathcal{P}_{F'} \leq n^{\beta(K-q)} \sum_{z^n \in A^n} 2^{-n\hat{H}_{F'}(z^n)} \,. \quad (17)$$

The *IMP-constrained type* of z^n relative to $F' = \mathcal{F}_k(\Pi', k)$ is the set of all conditional symbol counts of z^n for all the states of all the constituent Markov models of F' (these counts determine also the switch counts). It is readily verified that $q = Q(F')$ counts suffice to specify the type. Each term $2^{-n\hat{H}_{F'}(z^n)}$ in (17) depends on z^n only through its type, and, for each type, the terms added are probabilities of sequences assigned by the same distribution, which add-up to at most one. Therefore, the summation in (17) is upper-bounded by the number of types, which, in turn, is upper-bounded by n^{q+1} for sufficiently large n. Thus, from (17) we obtain

$$\mathcal{P}_{F'} \leq n^{q+1+\beta(K-q)}$$

The number of IFSMs with q parameters is very loosely upper-bounded by q^{q+1}. Recalling that $q \leq n$, we obtain

$$\sum_{Q(F') \geq Q_0} \mathcal{P}_{F'} \leq \sum_{q=Q_0}^{n} q^{q+1} n^{q+1+\beta(K-q)} \leq \sum_{q=Q_0}^{n} n^{2q+2+\beta(K-q)}$$
$$= \sum_{q=Q_0}^{n} n^{q(2-\beta)+\beta K+2} \leq n^{Q_0(2-\beta)+\beta K+3} \,, \quad (18)$$

where the last inequality follows from $\beta > 2$. Choosing $Q_0 > K + (2K+5)/(\beta-2)$, we obtain an error probability that decays faster than n^{-2}, and is thus summable. ∎

The results of this section extend, using similar tools, to the case where a different order is estimated for each Markov source, and also to the case where the interleaved sources are tree sources (see, e.g., [7]). In these cases, the estimator returns the estimated partition and the corresponding orders or trees.

V. Experimental results

We report on experiments that show the practical performance of the proposed deinterleaver. The table below shows deinterleaving success ratio for sequences of various lengths. For each length, 200 sequences were tested. Each sequence was generated by an IMP with m=3, sub-alphabet sizes α_1=4, α_2=5, α_3=6, Markov processes of order 1 with randomly chosen parameters, and uniform switch distribution. We compare results for an implementation of the scheme of [1], and two variants of the proposed scheme, labeled (a) and (b). Variant (a) implements (6) via exhaustive search. Since this is rather slow, variant (b) uses a randomized gradient descent-like heuristic, which is much faster, and achieves virtually the same performance (for shorter sequences, the heuristic sometimes finds the correct partition even when it is not the one that minimizes cost; this explains the slightly better performance compared to the exhaustive search). For the scheme of [1], tolerances were optimized for each sequence length with knowledge of the correct solution; a main threshold proportional to $\sqrt{\log n/n}$ seems to work well across the range tested. The table shows that the proposed scheme achieves better than 50% accuracy for sequences as short as n=1000, and close to 100% accuracy for $n \geq 2500$, whereas the scheme of [1], although fast, requires much longer sequences, having correctly deinterleaved just one sequence in 200 for n=5000.

n	[1]	(a)	(b)
500	0.000	0.020	0.045
1000	0.000	0.518	0.551
2500	0.000	0.980	0.990
5000	0.005	1.000	1.000
10000	0.030	1.000	1.000
50000	0.462	1.000	1.000
100000	0.774	1.000	1.000
500000	0.970	1.000	1.000

Acknowledgment

Thanks to Erik Ordentlich for stimulating discussions. The work of W. Szpankowski was partially supported and substantially done while visiting Hewlett-Packard Laboratories, Palo Alto, CA.

References

[1] T. Batu, S. Guha, and S. Kannan, "Inferring mixtures of Markov chains," in *COLT*, 2004, pp. 186–199.

[2] N. Landwehr, "Modeling interleaved hidden processes," in *ICML '08: Proceedings of the 25th international conference on Machine learning*. New York, NY, USA: ACM, 2008, pp. 520–527.

[3] G. Schwartz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[4] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[5] I. Csiszár and P. C. Shields, "The consistency of the BIC Markov order estimator," *Annals of Stat.*, vol. 28, no. 6, pp. 1601–1619, 2000.

[6] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258–1270, 1992.

[7] M. J. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Trans. Inform. Theory*, vol. 41, pp. 643–652, may 1995.

[8] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1007–1016, 2006.

[9] M. J. Weinberger and M. Feder, "Predictive stochastic complexity and model estimation for finite-state processes," *Journal of Statistical Planning and Inference*, vol. 39, pp. 353–372, 1994.